Original scientific paper
Received: September 5, 2022
Accepted: November 25, 2022

Lidija Nikolić, PhD, Assistant Professor
Josip Juraj Strossmayer University of Osijek
Faculty of Education
lnikolic@foozos.hr
https://orcid.org/0000-0003-0360-0469

VERIFICATION OF THE METRIC PROPERTIES OF E. E. GORDON'S ADVANCED MEASURES OF MUSIC AUDIATION AMONG NON-MUSIC MAJOR STUDENTS IN CROATIA

Abstract: *The aim of the study was to investigate the metric properties of the Advanced Measures of Music Audiation test by E. E. Gordon (1989) on a sample of nonmusic major students in Croatia. The study was conducted with female students of early and preschool education at the Faculty of Education in Osijek (N = 235). The descriptive indicators of the AMMA test and the normality of the distribution of scores in the Croatian sample do not differ from the results obtained during the standardization on the American sample and the Polish sample of nonmusic major students. It was found that the reliability coefficient of the AMMA test, calculated using the split-half technique, was 0.87, and high intercorrelation values were obtained between the Total test and the Tonal (r = 0.88) and Rhythm (r = 0.87) subtests. Lower values of the difficulty level and discrimination power (< 0.20) are the weakest features of the application of the AMMA test to the Croatian sample. Considering the overall observed metric properties, it can be concluded that the AMMA test is a reliable measurement tool for the assessment of auditory musical abilities in nonmusic major students in Croatia.*

Keywords: *music abilities, music aptitude, music education, measuring instrument, testing*

INTRODUCTION

The development of musical ability tests dates to Seashore and his first battery of tests in 1919. Since then, many different tests of musical abilities have been developed for diagnostic and predictive research and educational purposes (Larrouy-Maestri et al., 2019; Nikolić, 2017, 2022; Swaminathan et al., 2021). The intricacy of the construct of musical abilities complicates the selection of a measuring instrument for use in a particular educational or research situation and requires a very precise definition of the terms used by authors in describing each concept related to an individual's musical ability.

The conceptual definition of a person's ability to successfully engage in music is not clearly defined. A person's qualities to engage in music can be expressed by a plethora of terms (music ability, music abilities, musicality, music giftedness, music talent, music aptitude), which are often used as synonyms. This can lead to erroneous conclusions about this complex construct. The problem arises when a person's aptitude for learning music is equated with musical achievement. A person possessing music aptitude does not have to have musical training to prove his or her music aptitude via musical achievement. A child who has music aptitude does not have to want to be actively involved in music, nor do teachers and parents recognize him or her as such. Therefore, when discussing this phenomenon, it is necessary to determine what is meant by the terms used in the context of individual research. In this paper, the term *music aptitude* is used when referring to an individual's ability to learn music, which is not dependent on musical achievement and indicates the student's level of ability to learn music (Gordon, 2006). The term music *ability* refers to an individual's apperception of a musical component and ability to perceive and/or reproduce a specific musical component. The term *music abilities* will be used as a general term that encompasses music aptitude and all individual musical abilities, including musical achievement.

Music aptitude is the result of heredity and environment, but the extent of the influence of one and the other on an individual's music aptitude has not yet been explained (Gordon, 1989b). Music aptitude develops from birth to age nine, and its level represents aptitude for musical achievement and, like any other aptitude, exhibits a normal distribution in the population (Gordon, 1989b). After age nine, music aptitude stabilizes and remains the same throughout life (Gordon, 1999), determining the framework for an individual's maximum musical achievement (Gordon, 1989b).

Music education that enables the development of an individual's music aptitude and musical competence must begin in childhood, no later than age nine (Gordon, 2001; Radoš, 2010). However, music education can also begin later in life with the acquisition and development of musical competence. The assessment of a person's music aptitude at the beginning of music education is a valuable source of information for music educators who design the process of learning music by adapting procedures, methods, and techniques according to the indicators of the assessment of music aptitude of the person learning music. The music aptitude test is also required when accepting people into an ensemble where people of different ages can participate in various musical activities. The leader of the ensemble must assess the music abilities of the ensemble member to determine his/her role in the ensemble and adapt music literature and his/her way of working to the specifics of the participants' musical abilities.

Testing music aptitude or music ability is necessary for research that addresses the process of learning music, in which the individual's music aptitude is one of the most important factors. Music aptitude is an indispensable factor for musical achievement, but it is not the only one. In addition to musical abilities, an individual's musical achievement depends on numerous factors, such as personality traits, intellectual and sensorimotor abilities, volitional aspects, motivation for success in music, and numerous other psychological, social, economic, and music education factors that make learning music a multifactorial process. Add to the above factors the differences between certain forms of formal music learning in an educational context or informal forms of music learning in a quality recreational setting, as well as the different ages at which people learn music, and it becomes clear how music learning can be viewed from different perspectives. Researching the process of music learning, therefore, implies different measuring instruments that can be used to measure certain factors to reach scientifically sound findings about the process of achieving musical success, regardless of the level at which music learning takes place.

Given the complexity of the concept of musical abilities and the variability of forms of musical education and active music practice, as well as the need to explore an individual's musical abilities concerning their functioning in different areas of their lives, it is necessary to choose a reliable and valid measuring instrument for research. Of the numerous tests that have been developed to measure musical abilities, most are designed for children, while tests for adolescents and adults are less common. Gordon (1989b) claims that until 1989, there was no valid music aptitude test for undergraduate and graduate students. In Croatia, for music pedagogical purposes of music education of adolescents, music abilities are tested in a traditional way that lacks objectivity and reliability. For research purposes, the pedagogical Test of Musical Abilities (Nikolić, 2017) was designed for the student population, which is not intended for group testing and contains tasks related to music reproduction. This study aimed to verify the measuring instrument for testing the music aptitude of students who are just starting their education in music at university so that it can be used for research and music pedagogical purposes in Croatia, and it should meet several criteria:

a. the test should have validity, reliability, objectivity, discrimination power, and normativity
b. the test should be designed for adolescents and adults
c. the test should be validated on a representative sample
d. the tasks in the test should not be conditioned by the musical culture of the examinees
e. the test must be suitable for respondents with no prior musical training
f. the test should be auditory
g. the test should not be too long
h. the test should be suitable for group testing
i. the test should be available to researchers and use high-quality recordings.

The above criteria are met by Gordon's (1989a) Advanced Measures of Music Audiation test (*AMMA*).

ADVANCED MEASURES OF MUSIC AUDIATION (GORDON, 1989)

Gordon's research on musical abilities and testing them, as well as music learning, gave rise to his theory of audiation, music learning theory, and several tests of musical abilities for different ages following his theories (Gordon, 2011).

The term audiation was used by Gordon (1989b) to explain and refer to the process by which a person hears and understands music that is not physically present. Audiation implies the apperception of music, which means recognizing, identifying, and understanding musical content (*Hrvatska enciklopedija*, 2021). The process of audiation differs from imitation and memorization of music, which can occur without audiation, but in this case, they are not completely successful and are quickly forgotten (Gordon, 1989b). The understanding of notation also depends on audiation, and without it, improvisation or creation in music is not possible (Gordon, 1989b). Gordon (1989b) asserts that audiation is the foundation for the development and stabilization of music aptitude as well as all musical endeavors and achievement. For this reason, Gordon calls the *AMMA* test a test of music audiation, even though it is a test of music aptitude.

Gordon (2013) distinguishes between eight types of audiation that describe different thought processes that take place based on listening to music, such as giving syntactic meaning to music just heard that is no longer physically present (1), notational audiation (2, 3, 5, 7, and 8), remembering through audiation (4 and 5), and creating and improvising through audiation (6, 7, and 8). Audiation, according to Gordon (2013), occurs in six hierarchical and cumulative stages, each of which is the foundation of and combined with the next stage of audiation. In the first stage of audiation, short series of pitches and durations are recalled in the music just heard. If under conditions of immediate impression, the recalled series of pitches and durations are not given meaning, recall is lost. When the pitch and tone duration series are given meaning, the second phase of audiation can take place. In this phase, the retained tone sequences are imitated and organized into one or more tonal and rhythm patterns based on the pitch center and musical pulse. In the third stage of audiation, one becomes aware of the tonality and meter of the music heard because of the interaction of the tonal patterns and the mutual interactions of the rhythm patterns. In the fourth stage of audiation, tonal and rhythm patterns are retained based on the established tonality and meter so that in the fifth stage of audiation, we can recall and arrange them in other musical works we heard a few hours, days, or years ago, implying the establishment of similarities and differences between forms, which allows these processes to continue in the future. In the sixth stage of audiation, the tonal and rhythm patterns that will be heard in the continuation of the music heard are predicted, and better prediction leads to a better understanding of the music heard.

Based on his theory of audiation, Gordon developed music aptitude tests for preschool through college age. The Primary Measures of Music Audiation (PMMA) are designed for children ages five to eight who are in the developmental stage of developing music aptitude (Gordon, 2001). The Intermediate Measures of Music Audiation (IMMA) is designed for students ages six to nine who are in the transitional stage between the developmental and stabilized music aptitude stages and for students ages 10 and 11 who are in the stabilized music aptitude stage at ages 10 and 11 (Gordon, 1989b). The Musical Aptitude Profile (MAP) is intended for students who are in the stabilized music aptitude stage (ages 10 to 18). Gordon (2001) points out that it is particularly useful for diagnosing stronger and weaker musical qualities in each student.

The *AMMA* test was designed for use with high school students and undergraduate and graduate students (Gordon, 1989b) for career guidance, college admission, music conservatory, or ensemble admissions, designing effective music teaching, and setting objective and realistic expectations in music education work (Gordon, 1989b).

In designing the *AMMA* test, Gordon (1989b) followed several principles. The audiation of music was to be part of the test, not imitation, memorization, or discrimination of isolated pitches or tone lengths. The test is designed to be administered individually with one respondent or with a group of respondents so that it can be completed in a session of less than half an hour, and the test sheets should be suitable for manual and electronic processing of the results. A requirement that had to be met in the design of the *AMMA* test was that the test taker need not have knowledge of musical notation, music theory, or music history, nor need he or she be able to reproduce music by singing or instruments to be tested. The musical items in the test should be composed for the test to prevent the test taker from knowing the music. They should be performed by a professional musician, and the reproduction of the music in the test must be of the best technical quality attainable under practical conditions. The test should cover the full range of possible music aptitude and therefore be eclectic, i.e., include aspects of Gestalt and atomistic theories of musical abilities. The criteria of the items in the test included variability of tonality, meter, and tempo. Test takers should enjoy listening to the test, and listening to the test items should have an educational effect on them. The items in the test should not be arranged from easiest to hardest, but the order should be mixed to stimulate and maintain the respondent's attention. Furthermore, the nature of the answer should not be so complex that it requires other than musical ability, and the respondent should not be forced to answer if he/she is not sure, it is necessary to allow the respondent not to answer a certain task.

In contrast to the PMMA, IMMA, and MAP, which contain a separate rhythm and tonal subtest, so the test taker responds exclusively to a tonal or rhythm task, Gordon's (1989b) *AMMA* test is a more advanced form of the test in that there is no separate tonal and rhythm subtest. Thus, the tonal and rhythmic aspects are not heard separately, which is closer to the process of apperception of music, because the test taker hears tonality, keyality, melody, harmony, rhythm, meter, and tempo simultaneously. In each task, the test taker listens to the tonal and rhythmic aspects of the music task and chooses an answer from three options: tonal difference, rhythmic difference, and perceiving the same thing in both aspects.

AMMA contains 30 items programmed on a computer and played on an electronic instrument imitating the sound of a piano. Each task contains a short musical statement followed by a musical answer, and the respondent must decide whether the musical answer matches the musical statement. If the musical answer differs from the musical statement, the respondent must decide whether the difference is tonal or rhythmic. There may be one or two tonal or rhythm changes in musical answers, but differences in both properties never occur. To administer the *AMMA* test, test sheets were created with three columns (Same, Tonal difference, Rhythm difference), which the test taker uses to mark his or her answers. The test also contains three sample questions so that the test takers know what is required of them.

Of the 30 items in the test, 10 items have a tonal change, 10 items have a rhythm change, and in 10 items there is no change. Summing up the correct answers delivers a raw score that is not suitable for interpretation. Therefore, Gordon (1989b) adjusted the score to account for results in recognizing the same musical statements and those statements in which the test taker heard a difference in the answer to a musical statement that was not different, or he/she did not hear the difference. There is an advantage in the adjusted score – the test taker who left a blank answer has the upper hand in comparison to the one who solved it incorrectly after not being sure of the right answer. In this way, the obtained adjusted scores are useful for interpreting the overall test and particularly the tonal and rhythm subtests. Percentile ranks (Gordon, 1989b) were created for the three categories of respondents (high school students, music major students, and nonmusic major students), the reading of which can be used to assess an individual test taker's rhythm, tonal, and general musical aptitude. Music educators and researchers should be familiar with the construction of the test and the author's instructions in the manual supplied along with other materials in the Advanced Measures of Music Audiation – Complete Kit (Gordon, 1989a) to properly interpret and use the results of the *AMMA* test for diagnostic and prognostic purposes.

The *AMMA* test was standardized in the 1988/89 school year on a representative sample in the United States. The total number of respondents (high school students, music major students, and nonmusic major students) was 5336 (description of the sample can be found in Gordon, 1989b: 36–40). The reliability of the *AMMA* test was verified in the standardization process by the split-half technique (r = 0.81 – 0.88) and the retest (r = 0.83 – 0.89), conducted one week after the test, by calculating the standard error of measurement (2.6 – 3.7) and the standard error of a difference (1.7 – 2.0), making it a reliable test (Gordon, 1989b: 40–43). Reliability was also verified in relation to Gordon's MAP test, and correlation coefficients between related subtests ranged from 0.68 to 0.85 (Gordon, 1989b, p. 50). The validity of the *AMMA* test was verified by calculating the intercorrelations between the subtests and the Total test (0.72 – 0.95) and by calculating the item difficulty level (0.27 – 0.99) and discrimination power (0.20 – 0.69) (Gordon, 1989b). Gordon (2004) verified the reliability of the *AMMA* test on a sample of students aged 11 to 13 years (N = 2077) and calculated a high-reliability coefficient (0.80 – 0.86). Hanson (2019) analyzed 215 studies whose results were published in 47 journals and related to the use of Gordon's tests (*MAP, PMMA, IMMA*, and *AMMA*) and concluded that the tests were consistent, except that analysis of correlations between subtests showed mixed results.

Gordon himself (1997) emphasized the advantage of the *AMMA* test in scoring because the raw score does not reflect the true level of music aptitude. The correlations between the individual subtests and the Total test between the raw score and the adjusted score are very high (0.90 – 0.93) (Gordon, 1989b), indicating that adapting the raw score to the adjusted score does not affect the validity, but the adjusted score has better reliability. Gordon (1991) verified the reliability among German respondents (N = 129) and confirmed that the scoring procedure of the *AMMA* test should not be changed. The only weakness of the adjusted score in the *AMMA* test is that the adjusted score has high correlations between the Tonal and Rhythm subsets, which reduces the diagnostic value of the individual subtests (Altenmüller et al., 1997).

The study presented in this paper aimed to answer the question of whether Gordon's *AMMA* test is a reliable measuring instrument for assessing auditory musical ability in nonmusic major students in Croatia.

The aim of the study was to investigate the metric properties of E. E. Gordon's (1989a) Advanced Measures of Music Audiation test on a sample of nonmusic major students in Croatia.

The tasks of this study were as follows:

a. to determine the descriptive properties of the test
b. to confirm the reliability of the test
c. to determine the difficulty and discrimination power of the items in the test
d. to determine whether the metric properties of the test conducted on the Croatian sample of nonmusic major students are consistent with the metric properties of the *AMMA* test when standardized on the American sample
e. to determine how respondents feel about the *AMMA* test.

METHODOLOGY

Participants

The research participants were first-year female students in the Undergraduate University Study Program of Early and Preschool Education at the Faculty of Education in Osijek (N = 235). The average age of female students was 18.89 (SD = 0.823; Min. = 18; Max. = 22).

Procedure

The test was conducted at the beginning of the academic year during the first music classes. To include as many respondents as possible, the test was conducted at the beginning of five school years (from 2017/18 to 2021/22). Students were informed of the purpose of the survey and that the test would not be anonymous. Those who agreed to these conditions took the test.

The instructions for the *AMMA* test were translated from English to Croatian and replaced with the existing English instructions on the CD, as were the items on the test sheet (Gordon, 1989a). First, there were three sample items on the recording and the test sheet to adjust the volume before the test began and to make it clear to the respondents what exactly they needed to do. The meaning of the terms *tonal* and *rhythm* was explained to the respondents so that any possible misunderstanding of these terms would not affect solving the test. The test lasted 20 minutes, and the *AMMA* test was played on a CD stereo player with large speakers in a large lecture hall in groups of 20 students. Each respondent checked off the answer to the items on the test sheet after listening to each item, which consisted of playing two musical statements. In the 30 items of the *AMMA* test, each respondent could tick the answer in one of the three columns. In the first column, the respondent ticked if the two musical statements were the same, in the second column if there was a tonal difference in the repeated musical statement, and in the third column whether he/she heard a rhythmic difference in the repeated musical statement. Respondents were warned not to guess the answer if they were not sure but to leave the section blank for a given task.

After taking the *AMMA* test, students completed an anonymous questionnaire, *Evaluation of the AMMA test,* with four statements (*The instructions before the test were clear.*, *The tasks during the test were difficult.*, *I fully demonstrated my musical abilities.*, and *The test was not strenuous.*) with a five-point Likert scale of agreement (1 – *strongly agree*, 2 – *partially agree*, 3 – *neither agree nor disagree*, 4 – *partially disagree*, 5 – *strongly disagree*) and one with an open-ended question in which respondents could write in their own words what they thought about the method of testing musical abilities.

Data processing

Responses on the test forms were scored according to the instructions in the manual for conducting the *AMMA* test (Gordon, 1989b). The scoring procedure included three different outcomes. The highest possible score on the Tonal subtest is 40 points, the highest possible score on the Rhythm subtest is 40 points, and the highest possible score on the Total test is 80 points. The arithmetic mean (M), standard deviation (SD), standard error of measurement, reliability, and intercorrelation coefficient (r) were calculated according to the procedure described in the manual (Gordon, 1989b). The item difficulty level and discrimination power for 30 tasks of the *AMMA* test were also calculated. The survey was analyzed using descriptive statistics (M, SD) and content analysis. The statistical program SPSS was used for data processing.

RESULTS AND DISCUSSION

To determine whether the *AMMA* test is suitable for use with the students of early and preschool education, class teacher education, and other forms of music education of adolescents and adults in Croatia who are just beginning to learn music and have no experience with active music making or music education outside the general education (Nikolić, 2022), the results were compared with those of Gordon (1989b) in the test standardization process.

Descriptive characteristics of the *AMMA* test

The arithmetic mean and standard deviation were calculated for the Total test and each of the subtests (Tonal, Rhythm). By comparing the results with those reported by Gordon (1989b), it can be concluded that the scores of students of early and preschool education on the Total test and the subtests are somewhat lower than those of nonmusic majors and high school students in the United States (Table 1). The results in the sample of Croatian students are closer to the results obtained by Kołodziejski (2010b) with nonmusic major students in Poland (N = 552) (Table 1). Moreover, the results of another study by Kołodziejski (2010a) with future teachers (N = 50), nonmusic majors, are very similar to the Croatian sample of future early and preschool teachers. These comparisons show that the *AMMA* test is stable despite the cultural differences between respondents from the U.S., Poland, and Croatia.

Table 1

M, SD, Min., Max., skewness, and kurtosis on the AMMA test (Tonal and Rhythm subtest, Total test) for high school students and nonmusic majors[2]

AMMA	Early and preschool education students	USA nonmusic majors	USA high school students	Poland nonmusic majors**	Poland techer eduation

[2] The research results from Gordon (1989b) and Kołodziejski (2010a, 2010b) are taken with the number of decimals as stated in the literature.

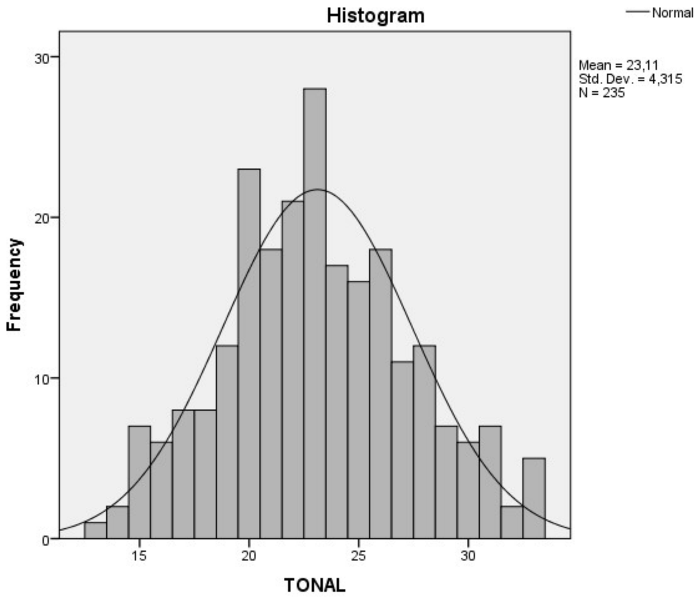
		Croatia N = 235	N = 2130*	N = 872*	N = 552	students*** N = 50
Tonal	M	23.11	24.30	23.80	23.05	23.90
	SD	4.32	4.89	4.37	3.9	3.65
	Min. – Max.	13 – 33				
	Skewness	0.138				
	Kurtosis	-0.349				
Rhythm	M	25.40	27.40	26.80	25.1	25.20
	SD	3.69	4.11	4.03	3.8	4.01
	Min. – Max.	15 – 36				
	Skewness	0.034				
	Kurtosis	0.055				
Total	M	48.43	51.70	50.60	48.1	49.00
	SD	7.57	8.49	7.91	7.3	7.46
	Min. – Max.	21 – 69				
	Skewness	0.066				
	Kurtosis	0.281				

* Gordon's results (1989b)
** Kołodziejski's results (2010b)
*** Kołodziejski's results (2010a)

To check the normality of the distribution of the results in the Total test and the subtests, histograms were constructed for each of them. Kurtosis and skewness were calculated as well. Histograms of the two subtests and the Total test (Graphs 1, 2, 3) show that the result values are normally distributed, which is also confirmed by the values for skewness and kurtosis (Table 1).

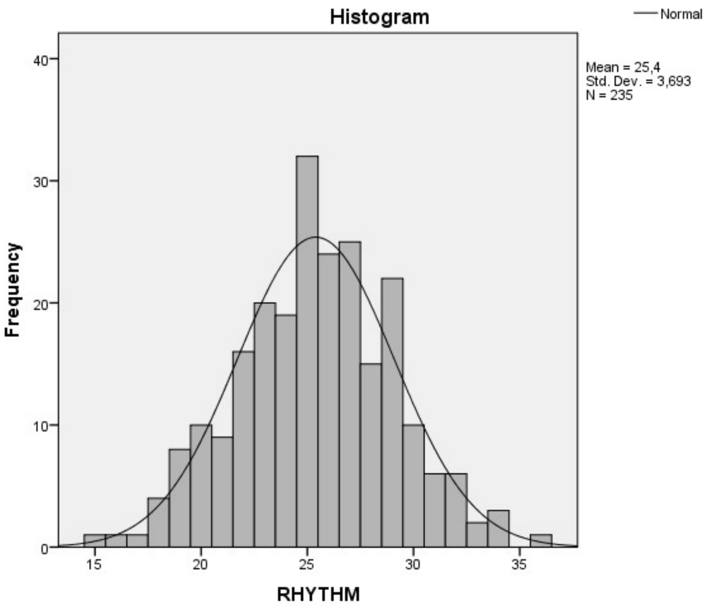
Graph 1

Tonal subtest histogram



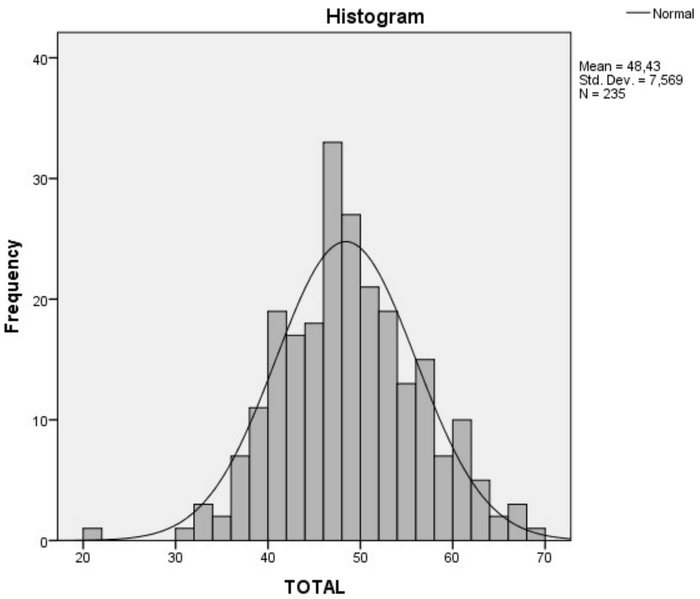
Graph 2

Rhythm subtest histogram



Graph 3

Total test histogram



Considering the established normality of the distribution of the results in the subtests and the Total test, as well as the descriptive indicators, we can conclude that these characteristics do not differ from the results obtained in the standardization process of the *AMMA* test.

Reliability

The reliability of the *AMMA* test on a Croatian sample of early and preschool teacher education students was tested using the split-half technique as described by Gordon (1989b) in the *AMMA* test standardization process. This refers to the method of selecting items for half of the subtests and the Total test, as well as to the calculation of the Spearman-Brown formula and the standard error of measurement (Bukvić, 2007). The results show moderate reliability of the Tonal and Rhythm subtests and high reliability of the Total test (Table 2). A comparison of the results shows that the coefficients in all three respondent groups are high according to Gordon (1989b: 40) and are above 0.80, while the reliability of the Rhythm subtest is moderate (*r* = 0.54) and that of the Tonal subtest is high (*r* = 0.72) for Croatian students. The Total test with Croatian early and preschool education students has high reliability (*r* = 0.87) and a small standard error of measurement. Compared to Gordon's test groups, the Rhythm subtest has lower reliability, which has a moderate reliability coefficient and a higher standard error of measurement. Compared to other relevant tests of musical abilities (as cited in Law & Zentner, 2012, p. 7), the *AMMA* also proved to be a reliable measuring instrument in the Croatian sample.

Table 2

The reliability coefficient (r) of the Tonal, Rhythm, and Total AMMA test

	Early and preschool education students Croatia (N = 235)			USA nonmusic majors (N = 2130) **			USA high school students (N = 872) **			USA music majors (N = 3206) **		
	TS*	RS*	TT*	TS	RS	TT	TS	RS	TT	TS	RS	TT
split-half	0.72	0.54	0.87	0.80	0.80	0.81	0.81	0.82	0.84	0.84	0.85	0.88
standard error of measurement	2.27	2.52	2.72	2.2	1.8	3.7	1.9	1.7	3.2	1.6	1.4	2.6

* Tonal subtest (TS), Rhythm subtest (RS), Total test (TT)

** Results are available in Gordon (1989b, p. 40).

By examining the intercorrelation coefficient (Table 3), we observed the extent to which the Tonal and Rhythm subtests of the *AMMA* test provide a unique measure of musical aptitude. In contrast to Gordon's results (1989b), in which the intercorrelation values were very high (0.91 – 0.95) between the Total test and the individual subtests, high (0.72 – 0.78) between the Tonal and Rhythm subtests for all respondent groups. This study found high intercorrelation values between the Total test and the subtests (0.87 – 0.88) and a moderate intercorrelation value between the Tonal and Rhythm subtests (0.66). Gordon (1989b) explains the high intercorrelations between the tonal and rhythm components of the *AMMA* test in part by tasks in the subtests in which there were neither tonal nor rhythm differences. The moderate value of the intercorrelation between the Tonal and Rhythm subtests in this study is justified considering that these are different musical abilities that need not be related, so this value is also significant. The values of all determined intercorrelations in the group of Croatian teacher education students are lower than the values determined by Gordon (1989b).

Table 3

Intercorrelation (r) of Tonal and Rhythm subtests, as well as Total test of AMMA

	Early and preschool education students Croatia (N = 235)			USA nonmusic majors (N = 2130) *			USA high school students (N = 872) *			USA music majors (N = 3206) *		
	TS	RS	TT	TS	RS	TT	TS	RS	TT	TS	RS	TT
Tonal subtest (TS)	-	0.66	0.88	-	0.72	0.93	-	0.74	0.95	-	0.78	0.94
Rhythm subtest (RS)	0.66	-	0.87	0.72	-	0.91	0.74	-	0.94	0.78	-	0.93
Total test (TT)	0.88	0.87	-	0.93	0.91	-	0.95	0.94	-	0.94	0.93	-

* Gordon's results (1989b).

In the abovementioned study by Kołodziejski (2010b), the intercorrelation for nonmusic majors between the Tonal and Rhythm subtests is 0.68, between the Tonal and Total test is 0.91, and between the Rhythm and Total test is 0.90, values very similar to Gordon's and the results of the present study.

From the consideration of two reliability measures, the reliability coefficient and the intercorrelation, it can be concluded that the *AMMA* test is a reliable measuring instrument. Due to its lower reliability, the Rhythm subtest is not suitable for independent use in the diagnosis of rhythm aptitude.

Item analysis

The item analysis in this study includes the calculation of the difficulty level and the discriminatory power of the individual items of the *AMMA* test. The item difficulty level indicates the percentage of respondents who answered a given item correctly. The item's discriminatory power describes the relationship between an individual respondent's answer to a single item and the overall score on the test. The higher the discriminatory power, the more effective the item is in distinguishing between participants who were successful in the Total test and those who were not. As seen in Table 4, the results show a wide range of values of item difficulty level ranging from 0.11 to 0.94 with an average value of 0.45, which is slightly below the optimal value of 0.5. A more detailed analysis shows that 18 items have a value of 0.20 to 0.60, while 6 items each have a value above 0.60 and below 0.20. In addition, the distribution of the item difficulty level was examined, and measures of skewness (0.343) and kurtosis (-0.312) of the distribution were calculated. The results show that the distribution is moderately and positively asymmetric, which means that a larger number of items have a value lower than the mean (0.45). A negative value of flattening indicates that the distribution is flatter than normal because there are more values at the edges of the distribution. However, these deviations do not significantly violate the normality of the distribution, which was confirmed by the Kolmogorov–Smirnov test ($Z = 0.095$; $p > 0.05$). Gordon's results (1989b) show an average difficulty level of 0.68, which is significantly higher than in this study. Analysis of item difficulty level in Gordon's study (1989b) shows that the range of values is from 0.27 to 0.99, with 14 items in the range of 0.20 to 0.60, 16 of which are above 0.60, and there are no items with values below 0.20.

.

Table 4

Difficulty level and discrimination power of items

Item	Task type*	Difficulty level**	Discrimination power
1	T	0.21	0.46
2	S	0.43	0.45
3	R	0.67	0.22
4	R	0.16	0.11
5	T	0.69	0.37
6	S	0.53	0.18
7	T	0.32	0.32
8	R	0.49	0.31
9	S	0.54	0.26
10	R	0.94	0.05
11	T	0.52	0.46
12	T	0.51	0.30
13	S	0.60	0.51
14	T	0.43	0.42
15	R	0.32	0.40
16	T	0.63	0.44
17	R	0.70	0.45
18	S	0.47	0.22
19	T	0.26	0.30
20	S	0.89	0.20

21	R	0.55	0.51
22	S	0.28	0.38
23	T	0.40	0.22
24	S	0.16	0.21
25	T	0.18	0.21
26	R	0.11	0.18
27	S	0.50	0.51
28	R	0.19	0.17
29	S	0.53	0.13
30	R	0.17	0.16

* T = tonal change; R = rhythm change; S = same musical statement.

** The average value of item difficulty level is 0.45, and the average value of discrimination power is 0.30.

From the comparison of the results, it can be concluded that the tasks of the *AMMA* test were more difficult for the students in Croatia than for the students in Gordon's study (1989b). Out of 6 items (4, 24, 25, 26, 28, 30) with a difficulty level lower than 0.20, five are included in the last 7 tasks of the test, which could indicate the respondents' fatigue in the last part of the test.

Discrimination power was determined by calculating the point-biserial correlation coefficient between the scores on a single item and the total score; more successful respondents are on top, and less successful respondents are on the bottom. The discriminatory power was calculated using the extreme group method (Husremović, 2016). Respondents were divided into three groups. The results of the group with lower scores on the test (0 – 11 correctly solved tasks; N = 67; 28.51%) and better scores on the test (16 – 30 correctly solved tasks; N = 66; 28.09%) were compared. The analysis of the values of discrimination power shows that all values are positive, 10 items have a value ≥ 0.40, which largely discriminates the sample in the *AMMA* test, 13 items have values between 0.20 and 0.39 and discriminate the sample to a lesser extent, while 7 items (4, 6, 10, 26, 28, 29, 30) with values below 0.20 discriminate the sample very poorly or not at all (Table 4). The average overall discrimination value is 0.30, which is lower than that in Gordon's study (1989b), where the average value of discrimination power is 0.40. In Gordon's research (1989b), there are 13 items with a value of discrimination power ≥ 0.40, and 17 of them have a value between 0.20 and 0.39, while there are no items with a value of < 0.20. Therefore, the *AMMA* test has weaker discrimination power in the sample of early and preschool education students in Croatia. When checking the discrimination power of the items, out of the 7 items that discriminate the sample poorly (< 0.20), 4 belong to the last 5 items on the test, which could indicate respondents' fatigue.

When comparing items with low difficulty level and discrimination power, items 4, 26, 28, and 30 were among the most difficult tasks and discriminated against the respondents poorly, indicating a possible difficulty. Considering that all the above items are rhythm tasks, this may explain the lower reliability of the Rhythm subtest shown earlier.

Thus, in addition to the weaker discrimination of the items of the *AMMA* test, it was also shown that the test items were more difficult for students in Croatia than for the sample in the standardization of the test. The weaker results in the last part of the test may indicate that the sample of students in Croatia had too many tasks and/or the test took too long, which should be verified by retesting, which was not done in this research.

Respondents' opinions about the *AMMA* test

Following the *AMMA* test procedure, a survey was conducted to determine respondents' opinions about this method of testing musical abilities and their experiences with the test. Respondents strongly agreed that the instructions prior to the test were clear (M = 1.08; SD = 0.337; N = 224; Min. = 1; Max. = 4). They expressed a neutral attitude toward having fully demonstrated their musical abilities (M = 3.14; SD = 0.977; N = 224; Min. = 1; Max. = 5). Partial agreement with the statement that the tasks were difficult (M = 1.99; SD = 0.961; N = 224; Min. = 1; Max. = 5) is not the best indicator of respondents' opinions, as there is a large dispersion for this result. Additionally, the result that respondents neither agree nor disagree with the statement that the test was not tiring (M

= 2.81; SD = 1.307; N = 224; Min. = 1; Max. = 5), with a large dispersion around the results, points to the different experiences of the respondents regarding the effort they felt during the test.

Due to the procedure of the *AMMA* test, it is not surprising that respondents were aware of how to solve the test and what the task was but also that due to the construction of the test, they could not assess whether they showed their best abilities. Respondent estimates of difficulty and effort varied considerably due to numerous causes, including their prior musical experience, experience with testing musical abilities and other cognitive abilities, and numerous other psychological causes.

Respondents could express their thoughts about the method for testing musical abilities. Since the question was optional, 38 of them provided the answer. Similar answers were counted, and the numbers of similar answers are provided after the statement in Table 5. The answers they wrote show that some of the respondents like this way of testing musical abilities because they find the test interesting. Responses that have a negative connotation regarding the experience with the *AMMA* test, in addition to unpleasant feelings, also describe the features of the test that bothered the respondents the most. The major objections to the test were that it contains too many tasks and that the time between musical statements is too short, which is why some of the test takers find the test tiring, complicated, and confusing, and find it takes too long.

Table 5

Respondents' opinions about the method of testing musical abilities (N=38)

If you wish, write down your opinion about the method of testing musical abilities.
It is interesting. (12)
I like this type of test; it requires deep concentration. (1)
I like this way of testing musical abilities. (1)
Terrible. (3)
It is difficult. (2)
It is very confusing. (2)
I don't like it; everything sounds the same and it is quite confusing and tiring. (1)
It is confusing when I cannot remember a musical statement, and then I'm not sure of the answer. (1)
I think the test takes too long and it is overtiring. (1)
I think there are too many musical statements and answers, so it is too tiring and complicated. (1)
I think there are too many items, and everything gets mixed up toward the end. (1)
There isn't enough time between two musical statements, so I cannot concentrate. (1)
There's too little time between the musical statements and there are too many of them. (1)
It is difficult to follow because the time between musical statements is too short. (1)
I was distracted by the announcements of the answers. (1)
Terrible, everything sounds the same after several items. (1)
Stressful. (1)
The testing method is interesting, but troubling because it seems very difficult. (2)
It is interesting, but it is difficult to distinguish and memorize musical statements. (1)
I like the method, but I was wondering all the time if I am extremely untalented. (1)
It is interesting but very confusing. (1)
It requires deep concentration and is quite challenging. (1)

Comparing the results of the item analysis with the comments of the respondents, the respondents explained why the items in the last part of the test showed greater difficulty level and weak discrimination power. The respondents found it difficult to solve that many test items, and they got tired in the last 5 – 7 items, which resulted in a lower score in these items, so they showed fewer differences between the respondents in terms of the tested trait of music aptitude.

Respondents indicated that it bothered them that they could not memorize musical statements because they changed quickly. Regarding the construction of the *AMMA* test, which was intended to create a test of music aptitude rather than musical achievement, it was necessary to disable memorization of musical statements to reduce the influence of music education on the measurement of music aptitude so that the test refers only to audiation, which requires approximately 4 seconds (Gordon, 1989b, p. 19), and a longer time allows for memorization.

Valerio et al. (2014) studied 112 amateur musicians with an average age of 67.8 years from the United States and Canada using the *AMMA* test and compared the arithmetic mean, standard deviation, reliability indicators, difficulty, and item discrimination power with the original study by Gordon (1989b) and concluded that this test is not a reliable measuring instrument for music aptitude for a sample of amateur musicians of this age. Indeed, the item difficulty level and discrimination power indicated that many items were too difficult for respondents and many of them discriminated poorly against the sample and, as indicated by respondents, the time between musical statements was too short to provide an answer (Valerio et al., 2014). The results of our study and those of Valerio et al. (2014) suggest the same weak features of the *AMMA* test, although respondents differ in age. The question arises whether the test should be revised to reduce the number of tasks. Considering that reducing the number of tasks could affect the reliability and discriminatory power of the test, a revision is not verified, as Valerio et al. (2014) also did not retest to confirm the data on the problematic items of the test. What a researcher using the *AMMA* test with a similar population could do to make the results of this instrument better indicators of actual music aptitude is to prepare the respondents for the procedure that awaits them by repeatedly conducting auditory exercises in which the same and different musical statements are perceived so that the respondents become accustomed to attention to listening during the *AMMA* test so that the results in the last part of the test are not weaker than in the rest of the test due to fatigue and lack of concentration. In addition, as in the study by Valerio et al. (2014), respondents complained that the time between musical statements was too short. Thus, they should be trained to solve the tasks faster because, at the time of the test, not all respondents may have the same experience with the timed form of the test. The suggested exercises could reduce the confusion that may hinder the examinee in performing the *AMMA* test.

CONCLUSION

An instrument designed to measure the influence of music aptitude on the process of learning music or other studies in which music aptitude is a possible factor must have satisfactory metric properties. To verify the metric properties of E. E. Gordon's *AMMA* test, a study was conducted with a sample of nonmusic major students in Croatia, but the process of acquiring their professional competencies implies the acquisition and development of musical competencies.

According to the descriptive indicators of the *AMMA* test and normality of distribution, the results of the Croatian sample do not differ from the results obtained when normalizing the American and Polish samples of nonmusic major students. It was found that the *AMMA* test has good reliability for measuring music aptitude of nonmusic majors in Croatia, as the reliability coefficient of the test is 0.87 and high intercorrelation values were obtained between the Total test and the Tonal (r = 0.88) and Rhythm (r = 0.87) subtests.

The analysis of the test items showed that the difficulty level of the items in the Croatian sample of nonmusic major students was lower (0.45) than that in the sample during the standardization process of the AMMA test (0.68). The average discrimination power is also lower in the Croatian sample of nonmusic majors (0.30) than in the American sample of nonmusic majors (0.40). The presence of items with difficulty level and values of discrimination power below 0.20 suggests that the metric properties of the AMMA test on the Croatian sample need to be further verified by a retest, which should show whether these indicators are caused by fatigue or whether there is another reason why certain items of the test are difficult for the Croatian respondents were more difficult and the respondents discriminated less well. Despite the lower average scores on the Total test and subtests, the normality of the distribution of scores was not disturbed.

Considering the totality of the examined metric properties in comparison with the previous results of Gordon and other researchers, it can be concluded that the *AMMA* test is a reliable measuring instrument for the assessment of auditory musical abilities in nonmusic major students in Croatia.

A limitation of the conducted research is the lack of a retest as a measure of reliability. Future research should therefore include a retest under the same testing conditions one week later, as was done in the standardization of the *AMMA* test (Gordon, 1989b). Future research, by administering a retest and analyzing the results, could account for the lower reliability coefficient within the Rhythm subtest when using the split-half technique but could also examine the rhythm items that had the lowest difficulty level and discrimination power.

The *AMMA* test, like other psychometric tests that measure only audiation musical abilities, is suitable for diagnostic purposes in research but not for prognostic purposes. The determination of higher levels of music aptitude, although reliable music aptitude tests such as Gordon's, cannot predict many desired outcomes of music learning (Hanson, 2019). Gordon (1989b) indicated that for prognostic music education purposes, it is best to combine a psychometric test with an examination conducted by an experienced music educator in the traditional, musical reproduction-based manner of the profession. Nikolić (2017, 2020) has shown how to construct a test of music ability administered by music educators that has all the good qualities of a test: validity, reliability, discrimination power, objectivity, and normativity. By using a reliable auditory test such as the *AMMA* test (Gordon, 1989b) in combination with a test that includes musical reproduction, such as the *Test of Musical Abilities* (Nikolić, 2017), a high level of diagnostic and predictive reliability can be achieved in assessing the musical abilities of individuals beginning education in adulthood, which should be verified by future research.

This study was an attempt to contribute to music pedagogical research and other research in which music aptitude and/or musical abilities are factors in the acquisition and development of musical competencies or are related to other human aptitudes, abilities, and functioning.

REFERENCES

Altenmüller, E., Gruhn, W., & Gordon, E. E. (1997). *Music, the brain and music learning: mental representation and changing cortical activation patterns through learning. Taking another look at the established procedure for scoring the advanced measures of music audiation*. Narberth (Pa.): Gordon Institute for Music Learning.

Bukvić, A. (2007*). Načela izrade psiholoških testova*. Zavod za udžbenike.

Gordon, E. E. (1989a). *Advanced Measures of Music Audiation - Complete Kit*. GIA.

Gordon, E. E. (1989b). *Manual for the Advanced Measures of Music Audiation*. GIA Publications.

Gordon, E. E. (1991). *The Advanced Measures of Music Audiation and the instrument timbre preference test: Three research studies*. GIA Publications.

Gordon, E. E. (1999). All about Audiation and Music Aptitudes: Edwin E. Gordon discusses using audiation and music aptitudes as teaching tools to allow students to reach their full music potential. *Music Educators Journal*, *86*(2), 41–44. https://doi.org/10.2307/3399589

Gordon, E. E. (2001). *Music Aptitude and Related Tests, an introduction*. GIA Publications, Inc.

Gordon, E. E. (2004). *Continuing studies in music aptitudes*. GIA Publications, Inc.

Gordon, E. E: (2006). Nature, source, measurement, and evaluation of music aptitudes. *Polskie Forum Psychologiczne, 11*(2), 227–237.

Gordon, E. E. (2011). *Roots of music learning theory and audiation*. GIA Publications.

Gordon, E. E. (2013). *Quick and Easy Introductions*. GIA Publications, Incorporated.

Hanson, J. (2019). Meta-analytic evidence of the criterion validity of Gordon's music aptitude tests in published music education research. *Journal of Research in Music Education, 67*(2), 193–213. https://doi.org/10.1177/0022429418819165

Hrvatska enciklopedija, mrežno izdanje (2021). Leksikografski zavod Miroslav Krleža. http://www.enciklopedija.hr/Natuknica.aspx?ID=3285

Husremović, Dž. (2016). *Osnove psihometrije: za studente psihologije*. Filozofski fakultet Univerziteta u Sarajevu.

Kołodziejski, M. (2010a). Musical aptitudes vs. readiness of pupils and students for harmonic and rhythm improvisation based on Polish research. *Problems in Music Pedagogy*, *7*, 47–65.

Kołodziejski, M. (2010b). Stabilized musical aptitudes as measured in Polish pedagogy students using Advanced Measures of Music Audiation test by Edwin E. Gordon. *Scholar Research Journal*, *13*, 8–17.

Law, L. N. C., & Zentner, M. (2012). Assessing musical abilities objectively: Construction and validation of the Profile of Music Perception Skills. *PLoS ONE, 7*(12), Article e52508. https://doi.org/10.1371/journal.pone.0052508

Larrouy-Maestri, P., Harrison, P. M. C., & Müllensiefen, D. (2019). The mistuning perception test: A new measurement instrument. *Behavior Research Methods*, *51*, 663–675. https://doi.org/10.3758/s13428-019-01225-1

Nikolić, L. (2017). Construction and validity of pedagogical test of music abilities. *Problems in Music Pedagogy*, *16*(2), 7–24.

Nikolić, L. (2022). Provjera metrijskih karakteristika Testa glazbenih sposobnosti. In A. Radočaj-Jerković i M. Milinović (eds.), *3. Međunarodni znanstveni i umjetnički simpozij o pedagogiji u umjetnosti - Inovativne metode poučavanja u umjetničkom području* (pp. 250–268). Akademija za umjetnost i kulturu Sveučilišta J. J. Strossmayera u Osijeku.

Radoš, K. (2010). *Psihologija muzike*. Zavod za udžbenike.

Swaminathan, S., Kragness, H. E., & Schellenberg, E. G. (2021). The Musical Ear Test: Norms and correlates from a large sample of Canadian undergraduates. *Behavior Research Methods, 53*(5), 2007–2024. https://doi.org/10.3758/s13428-020-01528-8

Valerio, W., Lane, J. S., & Williams, L. R. (2014). Using Advanced measures of Music Audiation among adult amateur instrumental musicians. *Research Perspectives in Music Education*, *16*(2), 2–15.